DEAR-Quest: Retrieval-Initialized Diagnostic State Tracking with Evidence-Grounded Question Expansion

Anonymous ACL submission

Abstract

Retrieval-augmented generation has become the standard approach for grounding large language models in external medical evidence, yet most pipelines treat retrieved content as static context for a single-step generation. Medical diagnosis is fundamentally different: it is a sequential, stateful process in which a clinician maintains competing hypotheses, acquires targeted evidence to discriminate among them, and updates beliefs as new information accumulates. We introduce **DEAR-Quest** (Diagnostic Evidence Attribution and Retrieval with Question Expansion), a retrieval-initialized diagnostic state tracking framework with a constrained attribution-grounded reader. DEAR-Quest first retrieves similar training cases to construct an explicit posterior over candidate diseases, then treats diagnosis as a multi-turn evidence-update process: each follow-up answer is converted into disease-specific support, contradiction, and uncertainty signals that revise the diagnostic state. To make this process effective, DEAR-Quest expands the question pool with train-derived clinical questions, explicitly models negative evidence for rule-out reasoning, and filters noisy question strings before they can affect the posterior. Finally, a constrained LLM reader consumes the completed posterior, question-answer trajectory, and attribution state, and must choose a final diagnosis from the candidate list. On the full DDXPlus benchmark (1,470 test cases), DEAR-Quest improves strict diagnostic accuracy from 70.95% (retrieval-only) to 83.61% with the non-LLM state tracker, and the constrained LLM reader further improves performance to 85.33%, while using 2.90 follow-up turns on average. Code is available at https://anonymous.4open.science/r/DEAR-2E06/.

1 Introduction

Large language models have made rapid progress on medical question answering and examination-style benchmarks (Singhal et al., 2023, 2025; Achiam et al., 2023; Nori et al., 2023), accelerated by instruction tuning (Li et al., 2023; Han et al., 2023; Bao et al., 2023) and retrieval-augmented generation (RAG), where external medical resources are supplied as grounding context (Xiong et al., 2024; Zakka et al., 2024; Wang et al., 2024). However, differential diagnosis is not simply a single-step answer selection problem over a complete vignette: a diagnostic system must form a working hypothesis space from partial evidence, decide what information to acquire next, and revise its belief state as new evidence supports or contradicts competing diagnoses.

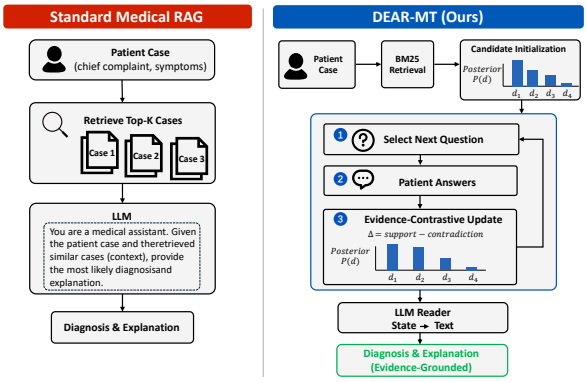

Figure 1: Comparison of standard medical RAG (left) and our DEAR-Quest framework (right). Unlike flat retrieve-then-generate pipelines, DEAR-Quest uses retrieved training cases to initialize an explicit diagnostic state, asks targeted follow-up questions, and updates the posterior with contrastive support and contradiction signals after each new evidence item.

This sequential nature exposes a limitation of standard medical RAG. Most systems retrieve documents or cases and then rely on the LLM's context window to implicitly maintain the diagnostic state. The retrieved evidence may improve factual grounding, but the posterior over candidate diseases, the rationale for selecting a follow-up

question, and the effect of each new answer are not represented as explicit objects. This makes it difficult to inspect whether a diagnosis changed because of clinically discriminative evidence, whether absent findings were used for rule-out reasoning, or whether the final answer merely reflects the LLM's latent preference among plausible disease names. For clinical decision support, the missing state representation is not a cosmetic issue: the system should expose what it believes, why it asks the next question, and how each answer changes the differential.

We propose **DEAR-Quest**, a diagnostic framework that treats retrieval as the beginning of an explicit state-tracking process rather than as context for immediate generation. Given the partial patient summary, DEAR-Quest retrieves similar training cases and converts their labels into an initial posterior over candidate diseases. It then expands the available action space with train-derived clinical questions, queries structured evidence through a controlled simulator, and updates the posterior with disease-specific support and contradiction signals.

The LLM component is deliberately placed after this non-parametric tracking process. Rather than allowing a free-form agent to retrieve, ask, and diagnose in one opaque loop, DEAR-Quest passes the completed posterior, the question-answer trajectory, and the attribution state to a constrained reader that must choose its final diagnosis from the top posterior candidates. This design keeps the diagnostic trajectory inspectable while still using an LLM for the final interpretation of structured state.

We evaluate DEAR-Quest primarily on DDXPlus (Fansi Tchango et al., 2022b), a benchmark with structured patient evidence and ground-truth diagnoses. Retrieval alone provides a 70.95% strict top-1 diagnostic state on the 1,470-case test split. The DEAR-Quest tracker improves this to 83.61% by acquiring targeted follow-up evidence and updating the posterior, and the constrained LLM reader further improves the final system to 85.33%. The main contribution is therefore a retrieval-initialized, multi-turn diagnostic state tracker that integrates clinical question expansion, negative-evidence modeling, attribution-aware posterior updates, and a candidate-constrained LLM reader in a single diagnostic pipeline.

2 Related Work

Retrieval-augmented generation grounds language models in external evidence by retrieving documents, passages, or memories before generation (Lewis et al., 2020; Guu et al., 2020; Karpukhin et al., 2020). Later work improves the retrieved context through multi-document fusion (Izacard and Grave, 2021), adaptive retrieval and self-critique (Asai et al., 2024), graph-structured organization (Edge et al., 2024), and retrieval-grounded reasoning (He et al., 2022). In medicine, RAG is commonly used to supply textbooks, guidelines, or domain corpora to an LLM before it answers a clinical question (Xiong et al., 2024; Zakka et al., 2024; Wang et al., 2024). Recent diagnostic RAG systems begin to move beyond single-turn prompting: MRD-RAG performs multi-round diagnostic retrieval and reasoning (Sun et al., 2025), while MEDDxAgent decomposes differential diagnosis into modular LLM agents (Rose et al., 2025). MedRAG further combines retrieved cases with structured diagnostic differences and proactive follow-up questions (Zhao et al., 2025). DEAR-Quest follows the same motivation of retrieval-grounded diagnosis, but uses retrieval to instantiate an explicit posterior state before the LLM reader is invoked.

Medical LLMs and diagnostic agents provide a second line of related work. Large models achieve strong performance on medical QA benchmarks (Singhal et al., 2023, 2025; Achiam et al., 2023; Nori et al., 2023), and medical instruction tuning improves domain adaptation in systems such as ChatDoctor, MedAlpaca, and DISC-MedLLM (Li et al., 2023; Han et al., 2023; Bao et al., 2023). Clinical agent benchmarks and systems further evaluate whether models can ask questions, interact with simulated patients, and refine a differential diagnosis over time, including AgentClinic (Schmidgall et al., 2024), AMIE (Tu et al., 2025), and MedAction (Hsu et al., 2026). These approaches demonstrate the promise of LLM-based interactive diagnosis, but much of the belief tracking remains embedded in generated text or latent model state. DEAR-Quest instead exposes the state as posterior probabilities and per-disease attribution scores, allowing the final LLM decision to be constrained by a structured diagnostic trajectory.

Automatic diagnosis research has also long studied symptom acquisition and sequential decision making. Earlier dialogue and diagnosis systems

model the interaction as reinforcement learning, hierarchical policy learning, or symptom sequence generation (Xu et al., 2019; Liao et al., 2020; Chen et al., 2022, 2023). DDXPlus and CASANDE provide a benchmark and decision-making framework for automatic differential diagnosis under partial evidence (Fansi Tchango et al., 2022b,a); later work such as AARLC, BED, and TPKA explores alternative policies and objectives for asking diagnostically useful questions (Yuan and Yu, 2024; Guan and Baral, 2021; Yang et al., 2024). DEAR-Quest is closest to this tradition in its multi-turn evidence acquisition, but differs in using retrieved training cases and empirical disease-conditioned statistics as a non-parametric diagnostic state tracker, with an LLM added only as a constrained reader of the completed state.

3 Task Formulation

Setting. Each patient case is associated with an initial patient summary x_0 containing demographic fields and an initial evidence item. The target is the ground-truth diagnosis $d^\star \in \mathcal{D}$. Each test case additionally contains structured clinical evidence fields $E = \{(f_j, v_j)\}$ covering symptoms and antecedents, but these fields are accessed only through an oracle structured patient simulator after the system selects a question. This setup isolates the diagnostic state-tracking problem: the system is evaluated on which evidence it chooses and how it updates belief, not on natural-language patient interviewing or symptom extraction.

Diagnostic state. At turn t, the system maintains a diagnostic state

$$S_t = \left\{ p_t(d),\ s_t(d),\ c_t(d),\ u_t(d) \right\}_{d \in \mathcal{C}_t}, \quad (1)$$

where $\mathcal{C}_t \subseteq \mathcal{D}$ is the current candidate set; $p_t(d)$ is the posterior probability of disease d; and $s_t(d)$, $c_t(d)$, $u_t(d)$ are per-disease support, contradiction, and uncertainty attributions accumulated through turn t.

Interaction loop. The system repeats three steps per turn: (1) select a feature $q_t \in \mathcal{Q}$ from the question pool using S_t; (2) query the oracle simulator to obtain $a_t = \mathrm{sim}(q_t, E)$; (3) compute evidence-contrastive attribution and update S_t. After $T_{\max}$ turns or when a stopping criterion is met, the system either returns the top-posterior disease from the tracker or passes the completed state to a constrained LLM reader that selects a final diagnosis from the candidate set.

Objective. The primary objective is strict top-1 accuracy: the predicted disease must exactly match the ground-truth pathology label. The LLM is constrained to select its final diagnosis from $\mathcal{C}_T$, bounding performance by candidate recall.

4 Method

Core idea. DEAR-Quest separates diagnosis into two coupled operations: *state construction* and *state revision*. First, it retrieves clinically similar training cases from the partial patient summary and converts their diagnoses into an explicit posterior over a small candidate set. This posterior is the diagnostic state, not merely retrieval context for generation. Second, DEAR-Quest repeatedly selects discriminative clinical questions, observes one additional evidence item, and updates the state by asking whether the new evidence supports each candidate more or less than the current posterior-weighted alternative. Clinical question expansion increases the evidence that can be acquired, whereas the evidence-contrastive update defines how each acquired item changes the diagnostic state.

4.1 Candidate Disease Initialization

DEAR-Quest builds a BM25 retrieval index (Robertson and Zaragoza, 2009) over the training split. The initial retrieval query is constructed solely from the information available before any follow-up question: $q_0 = \mathrm{textview}(x_0) = [\text{age, sex, initial evidence}]$. Full structured clinical evidence is intentionally excluded from q_0.

The retriever returns the k most similar training cases $\mathcal{R}(q_0) = \{r_1, \ldots, r_k\}$. Each retrieved case votes for its ground-truth disease, weighted by its BM25 score:

$$\mathrm{score}_0(d_i) = \sum_{r \in \mathcal{R}(q_0)} \mathrm{BM25}(q_0, r) \cdot \mathbf{1}[y_r = d_i]. \quad (2)$$

The initial posterior is obtained by normalizing these vote scores:

$$p_0(d_i) = \frac{\mathrm{score}_0(d_i)}{\sum_{d_j} \mathrm{score}_0(d_j)}. \quad (3)$$

The diseases with non-zero p_0 form the initial candidate set $\mathcal{C}_0$, which serves as both the retrieval-only prediction and the starting state for subsequent updates. Thus retrieval has two roles: it proposes the differential diagnosis and assigns the initial belief mass that later evidence can move.

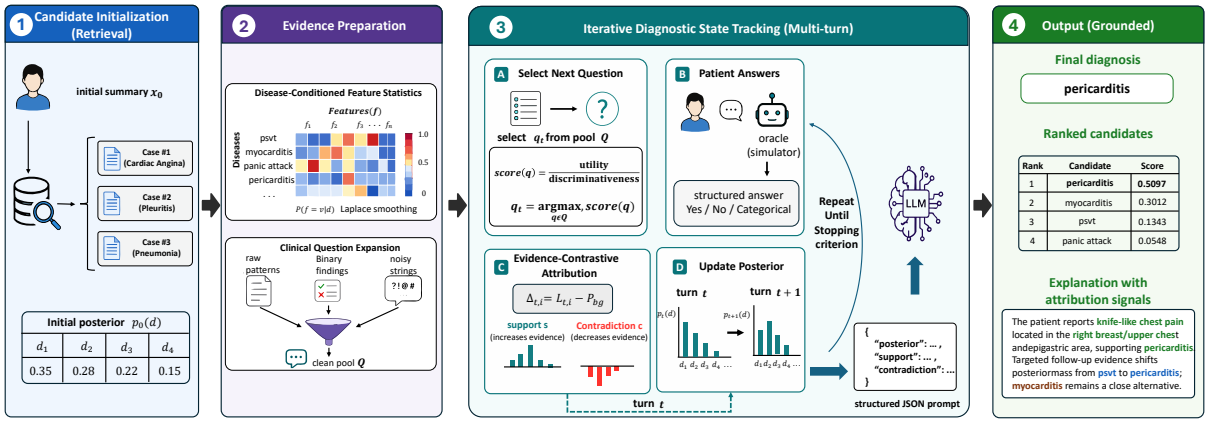

Figure 2: Overview of DEAR-Quest. The non-parametric state tracker uses retrieved training cases to initialize an explicit diagnostic posterior, expands the available evidence-acquisition actions with train-derived clinical questions, and revises the state after each follow-up answer through contrastive posterior updates.

4.2 Disease-Conditioned Feature Statistics

DEAR-Quest estimates empirical feature likelihoods from the training split using Laplace smoothing:

$$P(f{=}v \mid d) = \frac{N(f{=}v,\, d) + \alpha}{N(d) + \alpha\, |\mathcal{V}_f|}, \tag{4}$$

where $N(f{=}v, d)$ is the count of training cases with disease d and feature value v, $N(d)$ is the total count of disease d, $\mathcal{V}_f$ is the value set of feature f, and α is a smoothing constant. These statistics are used for posterior updating, discriminative question scoring, and attribution computation.

4.3 Clinical Question Expansion

The original structured feature pool contains only a small number of manually exposed follow-up fields, which causes multi-turn diagnosis to focus disproportionately on pain location and pain description. DEAR-Quest expands the action space by mining question-like clinical evidence patterns from the training split and converting them into askable features of the form

$$\texttt{question::}\langle\text{normalized clinical question}\rangle. \tag{5}$$

For example, training-set evidence such as whether the patient has shortness of breath, cough, fever, wheezing, or exertional symptoms is converted into candidate follow-up questions that can be selected during the diagnostic episode. These train-derived questions participate in the same empirical feature statistics as pain and demographic features, so they can be scored by pairwise discrimination, used in information-gain fallback, and applied to posterior updating.

For binary clinical questions, DEAR-Quest explicitly includes a FALSE value even when the training data mostly records positive findings. This negative-evidence model lets the system use absent findings as rule-out evidence: if a disease frequently presents with a finding, observing its absence contradicts that disease; if the finding is rarely associated with the disease, its absence should have little effect. Operationally, expanded binary questions use the simulator semantics in Section 4.5: a question recorded in a test case returns its observed value, an absent expanded binary finding returns FALSE, and non-binary unrecorded fields return UNKNOWN. Thus FALSE means absence in the structured DDXPlus schema, not an explicitly documented negative in free text.

We also apply a deterministic noise-aware filter before a mined string enters the action pool. The filter lowercases and normalizes whitespace, removes parser prefixes, and rejects answer-like fragments such as isolated letters, bare numbers or units, `na`, location/color remnants, and strings too short to be clinical predicates. This reduces artifact actions and value leakage; Appendix C.2 describes an excluded exploratory policy where residual strings still resembled feature-value probes.

4.4 Discriminative Question Selection

At each turn, DEAR-Quest selects from the structured question pool $\mathcal{Q}$. For two candidate diseases d_i and d_j, the discriminative score of question q is:

$$\mathrm{Disc}(q, d_i, d_j) = \max_{a \in \mathcal{V}_q} \big| P(a \mid q, d_i) - P(a \mid q, d_j) \big|. \tag{6}$$

The default policy selects the question most discriminative between the current top-2 posterior diseases:

$$q_t = \arg\max_{q \in \mathcal{Q} \setminus \mathcal{A}_t} \mathrm{Disc}(q, d_{(1)}, d_{(2)}), \tag{7}$$

where $\mathcal{A}_t$ is the set of already-asked questions. When pairwise discriminative features are unavailable, the policy falls back to expected information gain:

$$\mathrm{IG}(q) = H(p_t) - \mathbb{E}_a\big[H(p_{t+1} \mid q, a)\big], \tag{8}$$

where $H(p_t) = -\sum_i p_t(d_i) \log p_t(d_i)$. The predictive answer distribution is the posterior mixture

$$P(a \mid q, S_t) = \sum_{d_i \in \mathcal{C}_t} p_t(d_i)\, P(a \mid q, d_i). \tag{9}$$

For each possible value $a \in \mathcal{V}_q$, DEAR-Quest applies the same update rule hypothetically to obtain $p_{t+1}(d \mid q, a)$ and computes the expected posterior entropy under $P(a \mid q, S_t)$. This calculation uses the standard conditional-independence approximation that queried evidence fields are independent given the disease label; dependence is not modeled explicitly, but repeated questions are prevented by removing $\mathcal{A}_t$ from the pool.

Stopping rule. In all DDXPlus runs, inquiry stops when the posterior leader is sufficiently separated from the runner-up,

$$p_t(d_{(1)}) - p_t(d_{(2)}) \geq 0.30, \tag{10}$$

or when $T_{\max}$=5 is reached. Otherwise, the tracker returns the MAP disease at the final turn; sensitivity is reported in Appendix C.3.

4.5 Oracle Structured Patient Simulation

After selecting question q_t, the oracle simulator returns the value of the corresponding structured evidence field. If q_t is present in the case evidence E, it returns the recorded value v_{q_t}. If q_t is an expanded binary clinical question and the finding is absent, it returns FALSE; otherwise it returns UNKNOWN. This is a *controlled evaluation protocol*: the system may access a test-case field only after explicitly selecting it; initial retrieval does not use any structured evidence fields. The protocol therefore tests evidence selection and belief updating under partial observability, while abstracting away noisy natural language interaction. We refer to this as *structured-case-driven diagnostic simulation* throughout.

4.6 Evidence-Contrastive Attribution

For each evidence item (q_t, a_t) and each candidate d_i, we compute the disease-conditioned likelihood $L_{t,i} = P(a_t \mid q_t, d_i)$ from the pre-estimated statistics (Equation 4). We then compare it against a posterior-weighted background aggregating the competing candidates:

$$P_{\mathrm{bg}}(a_t \mid q_t, d_i) = \frac{\sum_{j \neq i} p_t(d_j)\, P(a_t \mid q_t, d_j)}{\sum_{j \neq i} p_t(d_j)}. \tag{11}$$

The contrastive evidence signal is:

$$\Delta_{t,i} = P(a_t \mid q_t, d_i) - P_{\mathrm{bg}}(a_t \mid q_t, d_i). \tag{12}$$

An alternative is a log-likelihood-ratio signal against the posterior mixture,

$$\Delta^{\log}_{t,i} = \log \frac{P(a_t \mid q_t, d_i)}{\sum_j p_t(d_j) P(a_t \mid q_t, d_j)}. \tag{13}$$

We use the probability-difference form in the main model because it is bounded in $[-1, 1]$ and is therefore less sensitive to rare disease–feature counts under empirical smoothing. The likelihood-only baseline in Section 6.3 isolates the effect of evidence acquisition without this contrastive background; we treat the log-ratio version as a principled robustness variant for calibration analysis. We decompose this into non-negative support and contradiction:

$$s_{t,i} = \max(\Delta_{t,i},\ 0), \tag{14}$$

$$c_{t,i} = \max(-\Delta_{t,i},\ 0). \tag{15}$$

A positive $\Delta_{t,i}$ indicates that the observed answer is more compatible with d_i than with competing candidates; a negative $\Delta_{t,i}$ indicates the opposite.

4.7 Posterior Update

The evidence-contrastive posterior update is:

$$p_{t+1}(d_i) \ \propto\ p_t(d_i) \cdot \exp\Big(\beta\big(w_s\, s_{t,i} - w_c\, c_{t,i}\big)\Big), \tag{16}$$

where w_s and w_c control support and contradiction influence and β controls update sharpness. The posterior is renormalized after each turn. We set w_u=0 in all main experiments (uncertainty excluded from the update; see ablation in Section 6.3).

4.8 Connection to Bayesian Belief Updating

The update in Equation 16 is a smooth approximation of Bayesian likelihood ratio updating. In exact Bayesian updating, $p_{t+1}(d_i) \propto p_t(d_i) \cdot P(a_t \mid$

$q_t, d_i)$; the likelihood $P(a_t \mid q_t, d_i)$ acts as an unnormalized weight. DEAR-Quest replaces the raw likelihood with its *contrastive residual* $\Delta_{t,i}$: the difference between the disease-specific likelihood and the posterior-weighted aggregate over competing diseases. This can be interpreted as a likelihood-ratio update where the denominator is the current marginal prediction of the evidence under the posterior mixture. When $\Delta_{t,i} > 0$, disease d_i explains the observation better than the field average, producing a positive multiplicative weight; when $\Delta_{t,i} < 0$, the disease is penalized. The exponential parameterization ensures positivity and differentiability of the update, and β controls the sharpness of the posterior shift per turn. Unlike standard likelihood updating—where a symptom common to all candidates contributes equal positive weight to every disease and therefore has little differential effect—the contrastive background term correctly identifies such evidence as non-discriminative ($\Delta_{t,i} \approx 0$) and leaves the posterior largely unchanged.

4.9 Attribution-Grounded LLM Diagnosis

After the non-LLM tracker finishes, DEAR-Quest can attach a constrained LLM reader. This reader is a post-hoc state reader, not an active question-selection agent: it does not perform new retrieval, ask additional questions, or modify the evidence-acquisition trajectory. It consumes the completed tracker output, including the initial patient summary, asked questions and simulator answers, final posterior, and the accumulated support, contradiction, and uncertainty attribution state.

The candidate list is the top-m diseases under the final posterior (m=5 in our main setting). The prompt instructs the LLM to choose exactly one final diagnosis from this list, to prefer the posterior top diagnosis by default, and to override it only when the question-answer evidence and attribution state clearly contradict the posterior top candidate. The LLM returns structured JSON with fields for the final diagnosis, ranked diagnoses, posterior top candidate, evidence support for the top candidate, largest contradiction, whether an override was allowed, and a concise reason. We then enforce the candidate constraint programmatically: if the LLM outputs a disease outside the candidate list, the prediction is mapped back to a valid candidate, with posterior top used as fallback. Thus the LLM reader functions as an attribution-grounded finalizer over an explicit diagnostic state rather than a free-form diagnostic agent.

5 Experimental Setup

5.1 Datasets

DDXPlus (primary). DDXPlus (Fansi Tchango et al., 2022b) is a large-scale synthetic benchmark for automatic medical diagnosis. Each case contains socio-demographic data, a ground-truth pathology, a differential diagnosis list, and structured binary, categorical, and multi-choice symptoms and antecedents. We use a local split of **11,760 training cases** and **1,470 test cases**. Structured evidence fields are accessed only through the oracle simulator; they are never used during initial retrieval.

Medical QA benchmarks. We also evaluate MedQA-USMLE (Jin et al., 2021), MedMCQA (Pal et al., 2022), and MMLU Medical (Hendrycks et al., 2020) as auxiliary fixed-question evaluation sets. These datasets do not expose a DDXPlus-style patient simulator or diagnostic posterior state, so they evaluate the retrieval-grounded reader interface rather than the full multi-turn diagnostic tracker.

5.2 Baselines and Variants

On DDXPlus we compare the final **DEAR-Quest** framework against initial BM25 retrieval, the non-LLM tracker without the reader, ablation variants of the tracker, and published DDXPlus systems. We treat the earlier DEAR-MT configuration as an ablation because the present framework changes the question-expansion, evidence update, and reader components. The final reader consumes the completed diagnostic state, ranked posterior candidates, and acquired evidence; it does not rerun retrieval or ask additional questions.

Published DDXPlus systems are included to contextualize performance, but they are not all protocol-matched baselines. Some report DDx-list metrics using complete differential information, whereas our main setting is an interactive top-1 diagnosis task with evidence acquired turn by turn. Accordingly, the direct internal comparisons are BM25 retrieval, the non-LLM tracker, the LLM reader variant, and controlled ablations under the same split and interaction budget.

For auxiliary generalization, we evaluate a lightweight QA-RAG adapter on MedQA-USMLE, MedMCQA, and MMLU Medical. This adapter retrieves solved training examples and constrains

Method	Protocol	Top-1 Acc.	Avg. Turns	Notes
AARLC (Yuan and Yu, 2024)	DDXPlus DDx list	75.39‡	25.75	RL-based
Diaformer (Chen et al., 2022)	DDXPlus DDx list	73.62‡	18.41	Transformer-based
CASANDE (Fansi Tchango et al., 2022a)	DDXPlus DDx list	69.70‡	19.71	RL + differential
Interactive top-1 systems				
MedRAG-GPT-4o (Zhao et al., 2025)	hierarchy level-3	73.23†	–	KG-elicited RAG
MEDDxAgent-GPT-4o (Rose et al., 2025)	interactive top-1	**86.00**	15.00	GPT-4o agent
BM25 retrieval	top-1 state	70.95	0.00	retrieval only
DEAR-Quest tracker	interactive top-1	83.61	2.90	non-LLM tracker
DEAR-Quest + LLM reader(GPT-4o)	interactive top-1	**85.33**	**2.90**	constrained reader

Table 1: DDXPlus diagnostic comparison. ‡GTPA@1 reported under the DDx list protocol where full patient evidence is available; not directly comparable to interactive top-1 where evidence is acquired turn by turn. †MedRAG reports DDXPlus level-3 accuracy under a hierarchy-specific protocol rather than GTPA@1. DEAR-Quest achieves 85.33% using only 2.90 average turns, approaching MEDDxAgent's 86.00% with one-fifth the queries.

decoding to answer options; it is reported separately because those benchmarks do not support the multi-turn diagnostic state tracker. Full protocol details are provided in Appendix B.2.

5.3 Metrics

The primary DDXPlus metric is strict top-1 accuracy. We also report average interaction turns and candidate recall hit@k. For QA benchmarks we report exact option or label accuracy.

5.4 Implementation Details and Hyperparameters

All DDXPlus runs use BM25 over the training split with k=20 retrieved cases and a fixed candidate set. Initial retrieval uses only age, sex, and the initial evidence field; remaining clinical evidence is accessed only through the simulator after a question is selected. The main configuration allows at most five follow-up turns, uses the pairwise-then-IG selection policy, enables negative-evidence updates, applies noise-aware question filtering, and sets β=1.5 in the update rule. The reader is evaluated with GPT-4o (Achiam et al., 2023) to match the backbone used by published baselines; DeepSeek-V4-Flash (DeepSeek-AI, 2026) results are reported in Appendix.

6 Results

6.1 Main Results

Table 1 shows that retrieval alone reaches 70.95% strict top-1 accuracy. Updating the retrieved posterior with interactive evidence raises accuracy to 83.61% using 2.90 turns on average. The constrained LLM reader further improves accuracy to 85.33%, approaching the 86.00% reported by MEDDxAgent-GPT-4o while using roughly one-fifth as many interaction turns. Relative to the non-LLM tracker, the reader converts 248 previously wrong cases to the correct diagnosis and changes 62 previously correct cases to incorrect predictions, for a net gain of 186 cases.

6.2 Generalization to Medical QA Benchmarks

Dataset	Items	Acc.
MedQA-USMLE	1,273	93.40
MedMCQA	4,183	80.73
MMLU-Med (overall)	1,089	94.12
Anatomy	147	92.59
Clinical Knowledge	265	92.08
College Biology	144	99.31
College Medicine	173	86.71
Medical Genetics	100	99.00
Professional Medicine	260	97.06

Table 2: QA-RAG adapter results on fixed-question medical benchmarks using BM25 top-3 retrieval and constrained option-letter decoding.

Table 2 reports auxiliary reader-interface results on fixed-question medical QA benchmarks. These results indicate that the retrieval-grounded reader can transfer to answer-option settings, but they should not be interpreted as evidence that the full diagnostic tracker transfers unchanged: the QA datasets lack a patient simulator, posterior state, and turn-level evidence acquisition. Because solved examples are retrieved from training pools, future releases should also include deduplication and train-test overlap audits for these auxiliary benchmarks.

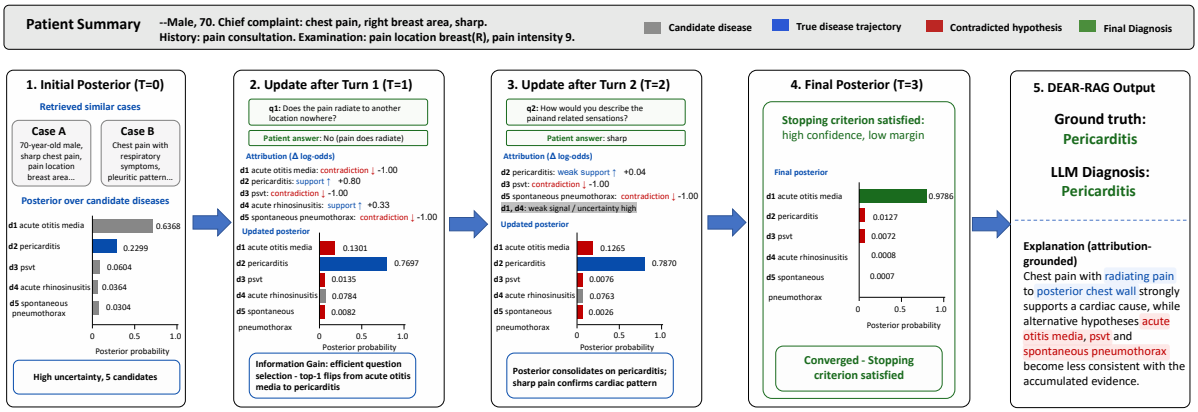

Figure 3: Qualitative case study of DEAR-Quest state updating on DDXPlus participant 1003. The initial posterior ranks *psvt* above the gold diagnosis, but targeted follow-up evidence shifts posterior mass toward *pericarditis*.

Variant	Acc.	Turns	Δ
DEAR-Quest full	**83.61**	2.90	—
w/o clinical question expansion	78.10	1.31	−5.51
w/o negative evidence modeling	78.78	3.36	−4.83
w/o noise-aware filtering	80.61	2.95	−2.99
Pairwise-only question policy	78.16	2.53	−5.44

Table 3: Ablation on the full DDXPlus test split (1,470 cases). Δ is absolute accuracy difference from DEAR-Quest full.

6.3 Update and Question-Policy Ablation

Table 3 isolates the main design choices. Removing clinical question expansion gives the largest drop (−5.51%), confirming that the initial DDXPlus question pool is too sparse for effective discrimination. Replacing pairwise+IG selection with pairwise-only scoring reduces accuracy by 5.44 points, showing that global posterior uncertainty matters beyond local disease-pair separation. Removing negative evidence reduces accuracy by 4.83 points, and disabling noise-aware filtering reduces accuracy by 2.99 points, indicating that both rule-out evidence and question quality affect the posterior trajectory.

6.4 Candidate Set and Retrieval Ceiling

Candidate set	Init. Acc.	Final Acc.	Hit@5	Avg. Turns
Top-3	76.26	79.93	88.57	1.12
Top-5	70.95	**83.61**	93.20	2.90
Top-8	70.95	75.99	95.85	1.44
Top-10	70.48	75.58	96.46	1.50

Table 4: Candidate-set size sensitivity. Top-5 best balances coverage and posterior discriminability.

Table 4 shows that the candidate set controls a coverage–discriminability tradeoff. Top-3 candidates give the strongest initial accuracy but miss more gold diagnoses. Top-8 and Top-10 improve candidate recall, but the larger candidate set makes posterior competition harder and reduces final accuracy. Top-5 gives the best balance, reaching 83.61% final accuracy with 93.20% hit@5. The constrained reader cannot recover diseases absent from the candidate set; this affects 6.8% of test cases and motivates future hybrid retrieval, reranking, or candidate-refresh mechanisms.

Appendix C.3 compares update families: the probability-difference update gives the best top-1 accuracy and Brier score, while log-ratio mixtures improve NLL and ECE at the cost of accuracy; DEAR-Quest degrades smoothly under answer flips and unknown-answer inflation, remaining above the retrieval-only baseline throughout. Appendix E shows that deleting acquired evidence changes the top diagnosis in 13.04% of turns and reduces confidence by 14.18% on average, though the correlation between attribution mass and confidence drop is modest (0.15), so attribution should be viewed as a diagnostic trace rather than a complete causal explanation.

7 Conclusion

We presented DEAR-Quest, a retrieval-initialized diagnostic state tracker with a constrained LLM reader. The non-LLM tracker improves strict DDXPlus accuracy from 70.95% to 83.61% via targeted evidence acquisition; the constrained reader further improves to 85.33%, with candidate recall remaining the clearest ceiling.

8 Limitations

DDXPlus is a synthetic benchmark generated from a rule-based medical knowledge base, so it does not capture the linguistic variability, noise, or incompleteness of real clinical encounters. Our simulator also returns exact structured field values; extending DEAR-Quest to free-form patient answers would require an additional language understanding layer. The negative-evidence mechanism is calibrated to this structured setting: FALSE means that a binary finding is absent from the case representation, not that a clinician explicitly documented a negative answer in free text. Likewise, mined question strings require dataset-specific leakage audits; the filtering in Section 4.3 reduces obvious artifacts but does not replace a release-time catalog of accepted and rejected patterns. Finally, the top-5 candidate set creates a hard ceiling: in 6.8% of test cases the correct disease is absent and cannot be recovered by the constrained LLM reader, motivating future work on candidate expansion and re-retrieval.

9 Ethical Considerations

DEAR-Quest is evaluated as a research system for diagnostic state-tracking, not as a deployed clinical decision-support tool. Its outputs should not be used for real patient care without clinical validation, human oversight, and evaluation on realistic patient interactions. Because DDXPlus is synthetic and structured, the reported results may not reflect performance on noisy free-form clinical dialogue. The constrained reader is designed to reduce unsupported free-form diagnoses by forcing selection from the candidate set, but this also means that retrieval failures can exclude the correct disease. The retrieval-initialized posterior may also transfer biases from the training cases, including demographic skews, and may encourage anchoring or premature closure when the first retrieved neighbors are misleading. Any clinical deployment would therefore need monitoring for candidate coverage, subgroup performance, override behavior, and attribution faithfulness, with clinicians retaining authority over diagnosis and follow-up.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pages 9112–9141.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv preprint arXiv:2308.14346*.

Junying Chen, Dongfang Li, Qingcai Chen, Wenxiu Zhou, and Xin Liu. 2022. Diaformer: Automatic diagnosis via symptoms sequence generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4432–4440.

Wei Chen, Cheng Zhong, Jiajie Peng, and Zhongyu Wei. 2023. Dxformer: a decoupled automatic diagnostic system based on decoder–encoder transformer with dense symptom representations. *Bioinformatics*, 39(1):btac744.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Arsene Fansi Tchango, Rishab Goel, Julien Martel, Zhi Wen, Gaetan Marceau Caron, and Joumana Ghosn. 2022a. Towards trustworthy automatic diagnosis systems by emulating doctors' reasoning with deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:24502–24515.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022b. Ddxplus: A new dataset for automatic medical diagnosis. *Advances in neural information processing systems*, 35:31306–31318.

Hong Guan and Chitta Baral. 2021. A bayesian approach for medical inquiry and disease inference in automated differential diagnosis. *arXiv preprint arXiv:2110.08393*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexei Figueroa, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hsin-Ling Hsu, Zizheng Wang, Donghua Zhang, Nai-Chia Chen, Jerry Wang, Jun-En Ding, Chia-Hsuan Hsu, Guoan Wang, Feng Liu, Fang-Ming Hung, and 1 others. 2026. Medaction: Towards active multi-turn clinical diagnostic llms. *arXiv preprint arXiv:2605.07305*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pages 874–880.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6769–6781.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Kangenbei Liao, CHENG ZHONG, Wei Chen, Qianlong Liu, Baolin Peng, Xuanjing Huang, and 1 others. 2020. Task-oriented dialogue system for automatic disease diagnosis via hierarchical reinforcement learning.

Seungseop Lim, Gibaeg Kim, Hyunkyung Lee, Wooseok Han, Jean Seo, Jaehyo Yoo, and Eunho Yang. 2025. H-ddx: A hierarchical evaluation framework for differential diagnosis. *arXiv preprint arXiv:2510.03700*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc.

Daniel Philip Rose, Chia-Chien Hung, Marco Lepri, Israa Alqassem, Kiril Gashteovski, and Carolin Lawrence. 2025. Meddxagent: A unified modular agent framework for explainable automatic differential diagnosis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13803–13826.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950.

Penglei Sun, Yixiang Chen, Xiang Li, and Xiaowen Chu. 2025. The multi-round diagnostic rag framework for emulating clinical reasoning. *arXiv preprint arXiv:2504.07724*.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, and 1 others. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450.

Yubo Wang, Xueguang Ma, and Wenhu Chen. 2024. Augmenting black-box llms with medical textbooks for biomedical question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1754–1770.

Cheng-Kuang Wu, Wei-Lin Chen, and Hsin-Hsi Chen. 2023. Large language models perform diagnostic reasoning.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end

knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7346–7353.

Yiqing Yang, Guoyin Zhang, Yanxia Wu, Zhixiang Zhao, and Yan Fu. 2024. Adaptive top-k algorithm for medical conversational diagnostic model. *Entropy*, 26(9):740.

Hongyi Yuan and Sheng Yu. 2024. Efficient symptom inquiring and diagnosis via adaptive alignment of reinforcement learning and classification. *Artificial Intelligence in Medicine*, 148:102748.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, and 1 others. 2024. Almanac—retrieval-augmented language models for clinical medicine. *Nejm ai*, 1(2):AIoa2300068.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, pages 4442–4457.

A Experimental Protocol Details

A.1 Dataset Splits and Task Formats

Dataset	Split	Train / retrieval pool	Eval. items	Task format
DDXPlus	test	11,760 training cases	1,470	multi-turn diagnostic state tracking
MedQA-USMLE	test	5,000 examples	1,273	multiple-choice medical QA
MedMCQA	dev	5,000 examples	4,183	multiple-choice medical QA
MMLU-Med	test	153 dev/validation examples	1,089	multiple-choice medical QA

Table 5: Evaluation protocols used in the main paper. DDXPlus is the only benchmark that directly supports the full diagnostic-state formulation. The other datasets are evaluated with the lightweight QA-RAG adapter described in Appendix B.2.

DDXPlus cases contain structured patient evidence, a diagnosis label, and disease hierarchy annotations. We use the training split as the case-retrieval pool and evaluate on the 1,470-case test split. The system first retrieves similar training cases, constructs a posterior over candidate diseases, asks structured follow-up questions, and updates the diagnostic state. The constrained LLM reader then receives the patient summary, asked Q/A evidence, final posterior, candidate list, and attribution state, and must select a final diagnosis from the candidate list.

MedQA-USMLE, MedMCQA, and MMLU-Med do not contain a DDXPlus-style disease posterior or structured patient simulator. We therefore evaluate them with a separate QA-RAG adapter: BM25 retrieves solved examples from the training or development pool, and the LLM answers the current multiple-choice question by exact option-letter matching.

A.2 Default Hyperparameters

Component	Default setting
Retriever	BM25 over training cases or solved QA examples
DDXPlus retrieval budget	top-20 retrieved cases
DDXPlus candidate cap	5 diseases
DEAR-Quest tracker maximum turns	5
LLM reader candidate cap	top-5 posterior diseases
LLM reader role	post-hoc constrained state reader; no new retrieval or questioning
QA-RAG retrieved examples	top-3 solved examples
LLM decoding	temperature 0
LLM output format	JSON with constrained labels

Table 6: Default implementation settings used by the reported runs.

B Prompting and Output Constraints

B.1 DDXPlus Constrained LLM Reader

The DDXPlus LLM reader is constrained to choose from the candidate diseases supplied by the non-parametric state tracker. The prompt contains five information blocks: patient initial summary, asked questions and answers, DEAR-Quest posterior, structured attribution state, and candidate diagnoses. The system instruction used by the reader is:

```
Use only the patient summary, asked Q/A evidence,
posterior state, attribution state, and candidate list.
Do not introduce diagnoses outside the candidate list.
Choose exactly one final diagnosis from the candidate list.
Prefer the posterior top diagnosis unless the Q/A
evidence and attribution clearly contradict it.
Return only JSON.
```

The user prompt asks the model to return a JSON object containing the final diagnosis, ranked diagnoses, posterior top candidate, evidence-supported top status, largest contradiction, whether the model is allowed to override the posterior, an override reason, and a short reason. If the model returns an invalid diagnosis, the implementation maps the answer back to the highest-ranked candidate rather than allowing an unconstrained free-form disease label.

B.2 Medical QA-RAG Adapter

For MedQA-USMLE, MedMCQA, and MMLU-Med, the reader adapter retrieves the top-3 most similar solved examples and asks the LLM to choose exactly one option. The system instruction is:

```
You are a medical exam question answering assistant.
Choose exactly one answer option from the provided
options.
Use retrieved examples as supporting context,
but answer the current question.
Return only JSON.
```

The user prompt contains the current question, answer options, retrieved examples, and the required JSON schema: `answer_idx`, `answer`, and `reason`. The `answer_idx` field is normalized to one of the available option letters before scoring.

Dataset	Eval. items	Accuracy	Fallbacks
MedQA-USMLE	1,273	93.40	0
MedMCQA	4,183	80.73	0
MMLU-Med	1,089	94.12	0

Table 7: Auxiliary fixed-question medical QA results. All rows use BM25 top-3 solved-example retrieval and constrained option-letter decoding.

MMLU-Med subject	Accuracy
Anatomy	92.59
Clinical Knowledge	92.08
College Biology	99.31
College Medicine	86.71
Medical Genetics	99.00
Professional Medicine	97.06

Table 8: Subject-level MMLU-Med accuracy for the updated reader run.

C Additional DDXPlus Experiments

C.1 Legacy Tracker Update and Question-Policy Ablation

Variant	Cases	Accuracy (%)	Δ vs. retrieval	Avg. turns
Initial retrieval only	1,470	70.95	—	0.00
1-turn update	1,470	76.26	+5.31	0.81
2-turn update	1,470	77.21	+6.26	1.28
Likelihood update only	1,470	77.07	+6.12	1.29
IG-only question policy	1,470	77.14	+6.19	1.27
Pairwise-only question policy	1,470	77.07	+6.12	0.97
Pairwise + IG default	1,470	77.21	+6.26	1.28
Support + contradiction update	1,470	**78.03**	**+7.07**	1.37

Table 9: Legacy DEAR-MT ablations before clinical question expansion. These rows support the w/o clinical question expansion baseline in Table 3; the final DEAR-Quest result is reported in the main paper.

C.2 Question-Policy Exploration and IG-Only Audit

We additionally ran an information-gain-only question policy under the expanded clinical-question space. This exploratory run reached 92.24% accuracy with 1.67 average turns, but we do not report it as a main ablation because manual inspection showed that some high-information question strings still contain answer-like values, such as rash color, radiation location, or swelling location remnants. Such strings can make the action less like a natural clinical question and more like a leaked feature-value probe. We therefore treat IG-only as a policy exploration that requires stricter question–value decomposition before it can be used as a fair comparison. The main DEAR-Quest result uses the more conservative pairwise-then-information-gain policy and the noise-aware filter described in Section 4.3.

C.3 Update Rule, Calibration, and Hyperparameters

Update rule	Acc.	NLL	Brier	ECE	Conf.	Turns
Likelihood-only	75.58	1.4547	0.3218	**0.0424**	0.7388	1.26
Probability difference	**77.55**	1.4391	**0.3035**	0.0840	0.7078	1.28
Log-ratio mixture	76.33	**1.4365**	0.3127	0.0457	0.7437	1.25

Table 10: Update-rule closure experiments on the full DDXPlus test split. Probability differences favor accuracy and Brier score, while log-ratio updates improve NLL and ECE.

Setting	Acc.	NLL	Brier	ECE	Turns
β=0.5	77.76	1.5449	0.3503	0.1588	1.40
β=1.0	**78.50**	1.4852	0.3198	0.1310	1.35
β=2.0	77.55	**1.4391**	**0.3035**	**0.0840**	1.28
α=0.1	77.35	**1.4315**	**0.3019**	**0.0726**	1.28
α=1.0	77.55	1.4391	0.3035	0.0840	1.28
α=5.0	**78.10**	1.4637	0.3132	0.1076	1.32
$T_{\max}$=1	76.39	1.4894	0.3301	0.0966	0.81
$T_{\max}$=2	**77.55**	**1.4391**	**0.3035**	**0.0840**	1.28
$T_{\max}$=3	**77.55**	**1.4391**	**0.3035**	**0.0840**	1.28

Table 11: Hyperparameter sensitivity under the probability-difference update. A second turn improves accuracy, while a third turn adds no benefit under the current stopping rule.

C.4 Robustness

Perturbation	Acc.	NLL	ECE	Conf.
None	77.55	1.4391	0.0840	0.7078
Flip 5%	76.94	1.4537	0.0827	0.7030
Flip 10%	76.39	1.4761	0.0814	0.6980
Flip 20%	75.24	1.5199	0.0729	0.6942
Unknown 10%	77.28	1.4538	0.0907	0.6984
Unknown 20%	76.46	1.4722	0.0914	0.6890
Unknown 30%	75.92	1.4909	0.0916	0.6772

Table 12: Robustness to synthetic simulator perturbations. Accuracy degrades smoothly under answer flips and unknown-value inflation while remaining above retrieval-only accuracy.

D Retrieval and Candidate Analysis

Metric	Initial	Final
Candidate hit@1	71.09	76.33
Candidate hit@2	85.44	86.94
Candidate hit@3	90.07	90.41
Candidate hit@5	93.20	93.20
Unique recall@5	96.94	
Unique recall@10	98.84	
Unique recall@20	100.00	

Table 13: Candidate recall and retrieval ceiling on the full DDXPlus test split. The constrained LLM cannot recover a disease absent from the candidate set.

Candidate set	Cases	Initial	Final	Avg. turns
Top-3	1,470	76.26	**79.93**	1.12
Top-5	1,470	70.95	83.61	2.90
Top-8	1,470	70.95	75.99	1.44
Top-10	1,470	70.48	75.58	1.50

Table 14: Candidate-set size sensitivity on the full DDXPlus test split. Smaller candidate sets improve discriminability, while larger sets increase coverage but make posterior ranking more diffuse.

E Attribution Faithfulness

Metric	Value
Prediction change after deletion	13.04
Mean confidence drop	14.18
Median confidence drop	6.24
Mean drop on correct cases	15.38
Corr(final support, drop)	0.1488

Table 15: Counterfactual deletion on the full DDXPlus test split. Values except correlation are percentages.

F Error Analysis and Case Study

F.1 Correction and Failure Types

System	Top-1 Acc.	Avg. Turns
DEAR-Quest + GPT-4o	85.33	2.90
DEAR-Quest + DeepSeek-V4-Flash	85.15	3.00

Table 16: LLM reader comparison on DDXPlus. GPT-4o is used in the main table to match published baselines.

System	Wrong → right	Right → wrong	Wrong → wrong	Candidate missing
Legacy DEAR-MT support+contradiction	112 (7.62)	8 (0.54)	315 (21.43)	51 (3.47)
DEAR-RAG w/ DeepSeek-V4-Flash	166 (11.29)	21 (1.43)	261 (17.76)	51 (3.47)

Table 17: Strict top-1 transition analysis on the 1,470-case DDXPlus test split. Values are counts with percentages in parentheses. A candidate-missing case is one in which the gold diagnosis is absent from the initial candidate disease set, making constrained final diagnosis impossible.

The dominant failure mode is not the update equation alone but the candidate bottleneck: when the gold disease is absent from the initial candidate set, neither DEAR-Quest nor a constrained LLM finalizer can output it. Among cases where the gold disease is available, remaining errors often involve diseases with similar initial symptoms or cases where follow-up answers are uninformative under the available structured features. The earlier DEAR-RAG finalizer corrected more initially wrong cases than the legacy tracker, but it also introduced more right-to-wrong transitions because the LLM finalizer may re-rank candidates based on retrieved examples and explanatory evidence.

F.2 Example Diagnostic Trajectory

G DEAR-Agent: Active Diagnostic Extension

DEAR-Agent extends the passive DEAR-MT tracker into an active inquiry loop by coupling the diagnostic state with a question-selection policy. At each turn t, DEAR-Agent selects the follow-up question that maximally discriminates among the currently plausible hypotheses:

$$q_t^* = \arg\max_{q \in \mathcal{Q}} \sum_{i<j} p_t(d_i)\, p_t(d_j) \cdot \mathrm{Disc}(q, d_i, d_j), \tag{17}$$

where $\mathrm{Disc}(q, d_i, d_j)$ is the expected information gain of question q for separating hypotheses d_i and d_j, approximated by the difference in BM25-weighted evidence attributions across the two candidate diagnostic clusters. The probability terms $p_t(d_i)$ and $p_t(d_j)$ act as importance weights so that the policy focuses on pairs that are still genuinely uncertain.

Stopping criterion. DEAR-Agent halts inquiry when the leading hypothesis achieves a sufficient margin over the second-best candidate:

$$p_T(d_{(1)}) - p_T(d_{(2)}) \;\geq\; \delta, \tag{18}$$

where $d_{(1)}$ and $d_{(2)}$ are the top-two hypotheses ranked by posterior probability at turn T, and δ is a threshold tuned on a held-out validation slice (we use $\delta = 0.30$ in all experiments). If the margin never exceeds δ within the maximum-turn budget $T_{\max}$, DEAR-Agent outputs the MAP hypothesis $d_{(1)}$ at $T_{\max}$.

Relation to DEAR-MT and DEAR-RAG. DEAR-MT is the passive tracker that updates the diagnostic state given whatever questions the environment provides; it forms the probabilistic backbone shared by all three variants. DEAR-RAG wraps DEAR-MT with a structured-state prompt and an LLM reader to produce free-text answers for open-domain QA benchmarks. DEAR-Agent additionally controls *which* question to ask next, making the full pipeline a closed-loop diagnostic agent suited to multi-turn clinical inquiry settings such as DDXPlus.

H External Direct LLM Results on DDXPlus

Recent work has reported direct LLM differential-diagnosis results on DDXPlus under evaluation protocols that differ from our strict top-1 multi-turn diagnostic-state setting. We include these numbers only as external reference points, not as directly comparable baselines. H-DDx evaluates LLMs on a 730-case stratified DDXPlus subset and asks models to generate a top-5 differential diagnosis list from the chief complaint and history of present illness (Lim et al., 2025). Table 18 reports representative DDXPlus results from that study.

Model	Top-5 Acc.	HDF1
Claude-Sonnet-4	83.90	36.73
Claude-Sonnet-3.7	83.60	33.80
Gemini-2.5-Flash	83.20	34.83
GPT-4o	80.40	34.99
GPT-4.1	80.10	33.87
GPT-5	78.30	34.48
Qwen3-235B-A22B	77.70	32.15
Qwen3-14B	77.50	33.67
MedGemma-27B	76.50	33.10
GPT-4o-mini	72.40	32.76
MediPhi	66.60	35.26

Table 18: External direct LLM DDx results on DDXPlus from H-DDx. These numbers use top-5 differential-diagnosis accuracy on a 730-case stratified subset, whereas our main DDXPlus evaluation reports strict top-1 accuracy on the full 1,470-case test split after multi-turn state updates.

An earlier diagnostic-reasoning study also used DDXPlus in a simulated doctor-patient dialogue setting with text-davinci-003 and small in-domain/out-of-domain subsets (Wu et al., 2023). Its reported accuracies are substantially lower than recent direct-LLM DDx results, so we treat it as related evidence for dialogue-style diagnostic prompting rather than as a primary comparison.

I Reproducibility Notes

All reported full-run results were written as JSON records with the configuration, per-example predictions, retrieved-example identifiers, and summary metrics. QA runs record the model name, base URL, retrieval top-k, random seed, worker count, and fallback count. For DDXPlus, each per-case record stores the initial posterior, selected follow-up questions, simulated answers, structured observations, attribution scores, updated posterior, and final prediction. This record format is what enables the transition analysis in Table 17.